Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 341689, 15 pages
doi:10.1155/2009/341689

Research Article

Multiuser Resource Allocation Maximizing the Perceived Quality

Andreas Saul and Gunther Auer

DOCOMO Euro-Labs, Landsberger Str. 312, 80687 Munich, Germany

Correspondence should be addressed to Andreas Saul, saul@docomolab-euro.com

Received 1 August 2008; Accepted 24 January 2009

Recommended by Thomas Michael Bohnert

Multiuser resource allocation for time/frequency slotted wireless communication systems is addressed. A framework for application driven cross-layer optimization (CLO) between the application (APP) layer and medium access control (MAC) layer is developed. The objective is to maximize the user-perceived quality by jointly optimizing the rate of the information bit-stream served by the APP layer and the adaptive resource assignment on the MAC layer. Assuming adaptive transmission with long-term channel state information at the transmitter (CSIT), we present a novel CLO algorithm that substantially reduces the amount of parameters to be exchanged between optimizer and layers. The proposed CLO framework supports user priorities where premium users perceive a superior service quality and have a higher chance to be served than ordinary users.

1. Introduction

With the high envisaged data rates of beyond 3rd generation (B3G) wireless communication systems [1, 2], multimedia broadband applications can be offered to mobile users. Multimedia applications are characterized by a multitude of data rate and quality of service (QoS) requirements. On the other hand, owing to the nature of the mobile radio channel, frequency selective fading, distance dependent path loss, and shadowing cause vast variations in the attainable spectral efficiency per user. The objective of multiuser resource allocation is to assign the available resources over the shared wireless medium to mobile users running different applications [3].

Orthogonal frequency division multiple access (OFDMA) provides orthogonal transmission slots in time and frequency, which may be flexibly assigned to the individual users [4, 5]. In B3G systems, this feature is exploited by the medium access control (MAC) layer to freely distribute the available bandwidth between users [6]. Provided channel state information at the transmitter (CSIT) is available, the number of transmitted information bits per slot can be adjusted to the channel conditions of a particular user.

The application (APP) layer outputs encoded applications, for example, a video stream. For the scalable video coding (SVC) extension [7, 8] of the advanced video coding (AVC) standard H.264/MPEG-4 AVC the stream may be received with a variable information bit rate. Other kinds of video streams may be encoded or transcoded [9] with the desired data rate. In general, any application may be delivered with variable information bit rate, allowing to trade user-perceived quality with data rate.

The high level of flexibility and adaptability offered by emerging system architectures provides an opportunity for dynamic allocation of resources across users and applications, to increase the network resource usage and to enhance the user satisfaction. This effectively requires interaction between system layers, a paradigm known as cross-layer design [10–12]. For the multiuser resource allocation problem at hand, a global cross-layer optimization (CLO) problem is formulated: maximize the user-perceived quality by tuning the served data rate on the APP layer jointly with the adaptive resource assignment on the MAC layer. Application-driven CLO has been studied for systems supporting one single type of applications [11, 13, 14] as well as for various application classes [15].

Several publications [15–17] consider a logarithmic relation between utility metric and data rate, which may result in a concave optimization problem. A more realistic utility metric, measuring the user-perceived quality, is given by the concept of mean opinion score (MOS) [18]. In [15],

a framework is established that allows to mathematically formulate the MOS for multiple applications like voice, video streaming, and file download. The resulting nonconcave optimization problem may be approximated, for example, with a greedy algorithm that maximizes the sum of the MOSs for all users [19].

In this paper, the optimum multiuser resource allocation supporting multiple applications is derived in closed form for the case of adaptive transmission with long-term CSIT, assuming a logarithmic relation between utility metric and data rate. Interestingly, the cross-layer optimization problem is shown to become independent of the channel conditions but is entirely determined by the application characteristics, provided that the offered data rate at the APP layer is matched to the adaptive transmission parameters in the MAC layer. For the special case where all users share the same application class, it turns out that the overall perceived quality is maximized when all users are allocated the same bandwidth, which corresponds to equal resource sharing. This implies that users with good channel conditions transmit with higher rate and therefore enjoy better QoS, as adaptive transmission is more bandwidth efficient in this case. This is in a sharp contrast to conventional approaches for QoS provisioning that assume a *fixed target rate* per user [3–5], where users with poor channel conditions are allocated more bandwidth, so that all receivers perceive the same QoS.

The theoretical analysis serves as a basis for a novel CLO algorithm that allows for a more realistic utility function that is based on the MOS. The proposed algorithm for the underlying nonconcave optimization problem is easy to implement and exhibits significantly lower complexity than the generic solutions in [19, 20]. Moreover, priority classes can be supported in the way that premium users perceive superior service quality and are more likely to be served, even under poor channel conditions. The proposed framework also allows to cater for additional constraints, such as a guaranteed minimum perceived quality for all users.

The developed CLO framework for application driven multiuser resource allocation is evaluated by mathematical and numerical analysis. We elaborate for which application classes CLO attains the most significant gains, and the origin of these gains is identified. Furthermore, the computational cost and the overhead due to exchange of CLO related parameters between layers is studied. It is demonstrated that the overhead of the proposed CLO framework grows only linearly with the number of users and available slots, which compares to an exponentially growing overhead for conventional techniques [11, 12, 21, 22]. This is particularly relevant to B3G systems with their high degree of freedom for resource allocation, due to the large number of served users and available slots.

The remainder of this paper is structured as follows. Section 2 provides an overview of the considered multiuser downlink with focus on MAC and APP layers. Section 3 introduces the CLO framework and the flow of exchanged parameters between layers and optimizer. In Section 4, the optimum multiuser resource allocation strategy is derived, assuming idealized application characteristics. The proposed CLO framework for the more realistic nonconcave optimization problem is established in Section 5, and its performance is evaluated by computer simulations in Section 6.

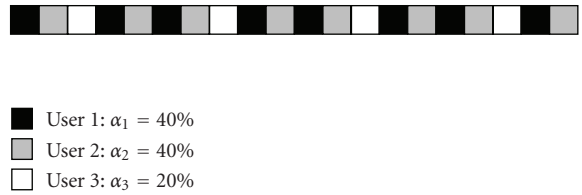

Figure 1: Packet-based generalized processor sharing (PGPS).

2. System Overview

A wireless downlink shared by K users is considered. An application server is transferring multimedia applications via core network and base station to mobile users. There are K applications, which, without loss of generality, generate K bit-streams, associated to K different users.

2.1. Link and Physical Layer. In the considered shared wireless downlink the resources are divided into slots occupying a given bandwidth and time, which can be flexibly allocated to users. A scenario where mobile users travel with potentially high velocities is considered. The high dynamics of the time varying channel prohibit the utilization of instantaneous CSIT. However, long-term CSIT that includes distance dependent path loss and log-normal shadowing is assumed to be available. As the long-term CSIT is constant over the whole frequency band, multiuser scheduling corresponds to the well known packet-based generalized processor sharing (PGPS) [23]. A PGPS scheduler aims to assign slots to user k proportionally to a coefficient α_k, which serves as input parameter for the scheduler, as illustrated in Figure 1.

The long-term CSIT allows to extract the average signal-to-noise ratio (SNR) for user k, which is used to select an appropriate modulation and coding scheme for the respective user. The spectral efficiency of the selected symbol mapping and coding scheme for user k is denoted by η_k in [bit/s/Hz]. Denote the number of symbols per slot by n_{slot}; the number of transmitted information bits per slot for user k amounts to $\eta_k n_{\text{slot}}$. Given user k is assigned all available slots N_{slot} exclusively, the maximum achievable data rate yields $R_{\max,k} = N_{\text{slot}} n_{\text{slot}} \eta_k$. The actual data rate to user k by the PGPS scheduler is then given by

$$R_k = \alpha_k R_{\max,k} = \alpha_k N_{\text{slot}} n_{\text{slot}} \eta_k. \tag{1a}$$

Additionally, the constraints

$$0 \le \alpha_k \le 1 \ \forall k \in \mathcal{K}, \quad \sum_{k \in \mathcal{K}} \alpha_k = 1 \tag{1b}$$

need to be fulfilled with $\mathcal{K} \triangleq \{1, \ldots, K\}$ being the set of all users; that is, the amount of assigned resources cannot be negative and the sum of all assigned resources equals the available resources.

2.2. Application Layer. The objective MOS is recommended as utility metric for voice transmission by the ITU-T [18] as a measure for the user satisfaction. Practically, the MOS may take values between 1 (not acceptable) and 4.5 (very satisfied). In [15], the MOS is extended to other services like video streaming, file download, and web browsing. The obtained mathematical model of the user-perceived quality can be used as universal utility metric for CLO, allowing for joint optimization of different application classes.

The application characteristic is mainly influenced by data rate and packet losses, described by the applications' rate-loss distortion [24]. In this paper, the perceived quality is exclusively expressed as a function of the data rate R_k, while packet losses are not considered as an explicit parameter. While this conveniently simplifies the analysis, this choice requires some further motivation, since certain kinds of source encoded bit-streams are sensitive to packet losses [11].

Packet losses may be caused by transmission errors over the mobile radio channel or by system overload. Regarding the wireless channel the link layer may compensate for packet losses by means of adaptive modulation and channel coding in combination with automatic repeat request (ARQ). While link adaptation ensures that transmission errors occur with low probability, low latency retransmissions of erroneous packets within the link layer [6] maintain reliable delivery of packets, at the expense of a certain rate reduction.

In an overloaded scenario, the offered load by the APP layer exceeds the capacity of the wireless link. Such an overload scenario can be effectively avoided by a fine grained adjustment of the offered data rate at the APP layer so as to match the capacity of the wireless link.

For instance, in case of video streaming, transcoding [9] or using the SVC extension of H.264/MPEG-4 AVC [7, 8] allows to vary the data rate in a rather fine granularity. As packets can be dropped at either the application server or the base station, a low latency rate adaption mechanism is feasible, at the same physical location as the scheduler in the MAC layer, effectively allowing to express perceived quality by data rate.

Moreover, the possibility to selectively drop packets offers one further opportunity to adjust the data rate. Likewise, for file downloads the data rate can also be adjusted in arbitrarily small steps. Hence, it is reasonable to assume that the application data rates can be adjusted continuously.

2.2.1. Video Streaming. We choose video streaming as one relevant example of an application class. In [25], a simple concave rate-distortion model is proposed for H.264/MPEG-4 AVC that relates the data rate of a video stream to the peak signal-to-noise ratio (PSNR):

$$\mathrm{PSNR}_k\big|_{\mathrm{dB}} = a + b\sqrt{\frac{R_k}{c}}\left(1 - \frac{c}{R_k}\right). \tag{2}$$

The parameters a, b, and c characterize a specific video stream or sequence, which is source encoded with rate R_k. These parameters may be determined by matching the distortion-rate model to the measured bit stream of a video.

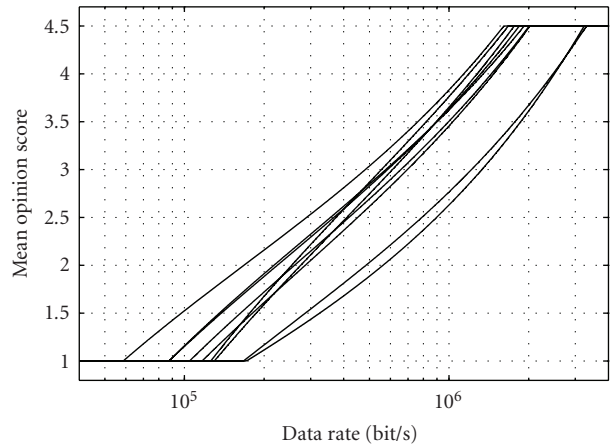

Figure 2: Time variant application characteristic of "Foreman" video stream.

According to [15, 26], the relationship between PSNR and MOS may be approximated by the bounded logarithmic function:

$$\mathrm{MOS}_k(\mathrm{PSNR}_k) = \begin{cases} 1 & : \mathrm{PSNR}_k \le \mathrm{PSNR}_{1.0}, \\ d\log\mathrm{PSNR}_k + e & : \mathrm{PSNR}_{1.0} < \mathrm{PSNR}_k < \mathrm{PSNR}_{4.5}, \\ 4.5 & : \mathrm{PSNR}_k \ge \mathrm{PSNR}_{4.5}, \end{cases} \tag{3a}$$

with

$$d = \frac{3.5}{\log\mathrm{PSNR}_{4.5} - \log\mathrm{PSNR}_{1.0}}, \qquad e = \frac{\log\mathrm{PSNR}_{4.5} - 4.5\,\log\mathrm{PSNR}_{1.0}}{\log\mathrm{PSNR}_{4.5} - \log\mathrm{PSNR}_{1.0}}. \tag{3b}$$

The parameters $\mathrm{PSNR}_{1.0}$ and $\mathrm{PSNR}_{4.5}$ denote the PSNR at which the perceived quality drops to "not acceptable" (MOS = 1.0) and exceeds "very satisfied" (MOS = 4.5), respectively.

The rate-distortion characteristic of a video typically varies over time, which means that the parameters a, b, and c are time variant. For example, during a scene cut a higher data rate is required to maintain a certain quality. As an example Figure 2 shows the rate-MOS model for $\mathrm{PSNR}_{1.0}$ = 30 dB and $\mathrm{PSNR}_{4.5}$ = 42 dB of the well known "Foreman" video. The 9 different curves correspond to different parts of the video of 1 second duration each.

3. Application-Driven Cross-Layer Optimization

Cross-layer design implies that additional parameters are to be exchanged between link and APP layers, denoted as control information. Figure 3 illustrates the system architecture including the flow of control information. In the following, the architecture, functional blocks, and variables depicted in Figure 3 are described.

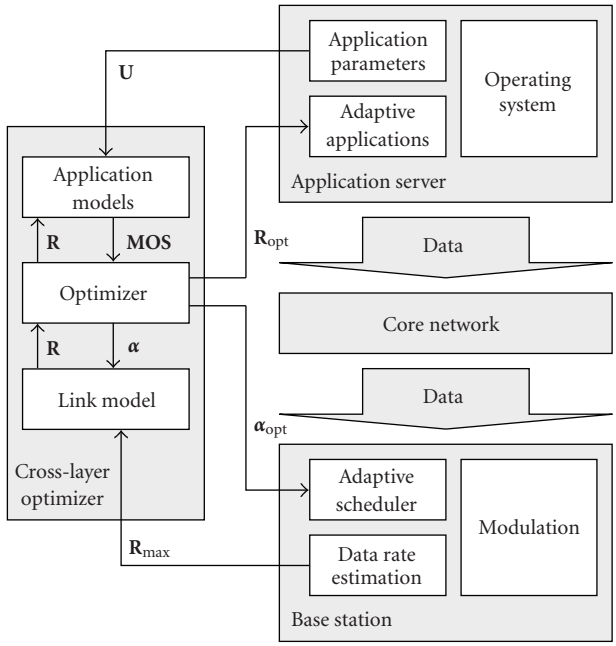

Figure 3: Control information processing and flow.

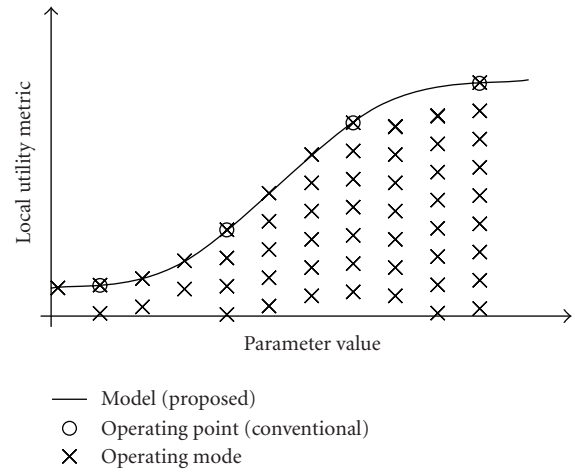

Figure 4: Visualization of operating modes.

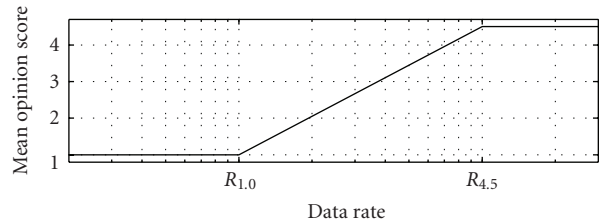

Figure 5: Considered generic application characteristic for one example application class.

3.1. Layer Model. A major challenge in cross-layer design is the abstraction of parameters exchanged as control information. In order to limit the amount of control information, we introduce a layer model at the optimizer that emulates the relevant characteristics of the layer. The parameters of the layer model are determined at the corresponding layer, and only these parameters are sent as control information to the optimizer. The optimizer then tunes the model so as to identify the operating modes that maximize the chosen utility, which are then fed back to the system layers.

Figure 4 demonstrates the difference between the proposed model-based approach, and conventional parameter abstraction based on operating modes (crosses) and points (circles) [11, 12, 21, 22]. The X-axis indicates the choice of one parameter a_1, and the Y-axis indicates the corresponding utility metric $u = \tilde{f}(a_1, a_2, \ldots)$. Depending on the choice of a_1 and further parameters $a_2, \ldots$ that cannot be determined from Figure 4 different operating modes of the utility metric are achieved.

For instance, applied to a video stream the local utility $\tilde{f}$ could be the PSNR or MOS, and according to (2) the parameters $a_1, \ldots$ might represent source coding parameters such as the chosen codec, the frame rate, and the data rate R_k. As a second example, applied to the PHY layer the local utility might be the sum throughput of all users, and $a_1, \ldots$ are parameters such as the channel coefficients or the velocity of the mobile terminal.

Following the conventional idea of parameter exchange, an intralayer optimization might deliver the subset of operating modes that maximize the utility function u, called efficient set in [22], also known as Pareto frontier. These operating modes are the crosses being located on the curve in Figure 4. A subset of operating modes is selected as operating points (circles). These are provided to the optimizer, which performs CLO by choosing the overall best operating point.

The proposed layer model is the curve in Figure 4, which represents an approximation of the utility metric $u = \tilde{f}(a_1, a_2, \ldots)$ as a continuous function. As demonstrated in the following the proposed parameter abstraction by a layer model exhibits a significant advantage for multiuser resource allocation, due to the potentially large number of available slots.

3.1.1. Link Layer Model. For conventional CLO the parameters that are provided to the optimizer are the set of possible data rates for all users $\{R_k\}$ in (1). Considering an OFDMA-based B3G air interface with a large number of available slots, a prohibitive set of possible data rates is obtained. Instead of offering a set of discrete values to the optimizer, the link layer model defines the shares of the available resources per users, $\alpha_k \in [0, 1]$ in (1), as continuous functions. The factors α_k allow the optimizer to tune the link layer model. Then, according to (1) an arbitrary number of data rate combinations $R_1, \ldots, R_K$ can be emulated at the optimizer. The only required parameters at the optimizer are the set of K parameters $\{R_{\max,k}\}$. Hence, the link layer model for the optimizer is fully determined by (1). Once the optimizer has found an optimum set of coefficients $\{\alpha_{\text{opt},k}\}$, these are fed back to the link layer.

3.1.2. Application Layer Model. The considered generic application characteristic resembles a bounded logarithmic relation between perceived quality and data rate as illustrated in Figure 5, described by the MOS as a function of the data rate R_k of user $k \in \mathcal{K}$

$$\mathrm{MOS}_k(R_k) = \begin{cases} 1 & : R_k \leq R_{1.0,k}, \\ \mathrm{MOS}_{0,k} \log \dfrac{R_k}{R_{0,k}} & : R_{1.0,k} < R_k < R_{4.5,k}, \\ 4.5 & : R_k \geq R_{4.5,k}, \end{cases} \tag{4a}$$

with

$$\mathrm{MOS}_{0,k} = \frac{3.5}{\log\left(R_{4.5,k}/R_{1.0,k}\right)}, \tag{4b}$$

$$R_{0,k} = R_{1.0,k}\left(\frac{R_{1.0,k}}{R_{4.5,k}}\right)^{1/3.5}, \tag{4c}$$

$$0 \leq R_{1.0,k} < R_{4.5,k} \quad \forall k \in \mathcal{K}. \tag{4d}$$

The semilogarithmic plot of Figure 5 visualizes the related parameters: the parameter $\mathrm{MOS}_{0,k}$ determines the slope of $\mathrm{MOS}_k(R_k)$ while $R_{0,k}$ shifts the curve along the X-axis.

Each user's application characteristic can be parametrized by only two parameters, $\{R_{1.0,k}, R_{4.5,k}\}$, or alternatively $\{\mathrm{MOS}_{0,k}, R_{0,k}\}$. The optimizer then tunes the model by maximizing the user-perceived quality and returns the set of optimum user data rates to the APP layer.

3.2. Parameter Exchange

3.2.1. System Description. Figure 3 shows a block diagram of the considered CLO framework and illustrates the signal flow of the exchanged control information between optimizer and layers. In order to formally describe the proposed model-based method of parameter exchange and optimization, we define the vector

$$\mathbf{R_{max}} \triangleq \left(R_{\mathrm{max},1}, \ldots, R_{\mathrm{max},K}\right)^{\mathrm{T}} \tag{5}$$

containing the maximum data rates of all users, the vector

$$\boldsymbol{\alpha} \triangleq \left(\alpha_1, \ldots, \alpha_K\right)^{\mathrm{T}} \tag{6}$$

containing the optimization coefficients, the vector

$$\mathbf{R} \triangleq \left(R_1, \ldots, R_K\right)^{\mathrm{T}} \tag{7}$$

containing the actual data rates of all users, and the vector

$$\mathbf{U} \triangleq \left(U_1, \ldots, U_K\right)^{\mathrm{T}}. \tag{8}$$

The parameter U_k describes the application characteristic for user k, which is $R_{1.0,k}$ and $R_{4.5,k}$ for the APP layer model from Section 3.1.2. In addition more detailed information about the applications in a real system may also be contained in U_k.

The link layer model described in Section 3.1.1 is defined by the vector function $\mathbf{f}_{\mathrm{L}} \triangleq \left(f_{\mathrm{L},1}, \ldots, f_{\mathrm{L},K}\right)^{\mathrm{T}}$ with elements

$$f_{\mathrm{L},k} : \alpha_k, R_{\mathrm{max},k} \longrightarrow R_k = f_{\mathrm{L},k}(\alpha_k), \tag{9}$$

which is given by (1). This means that based on the optimization coefficients $\boldsymbol{\alpha}$, which reflect the resource allocation on the link layer, the achievable data rates $\mathbf{R}$ of the users are determined.

The application layer models detailed in Section 3.1.2, $\mathbf{f}_{\mathrm{A}} \triangleq \left(f_{\mathrm{A},1}, \ldots, f_{\mathrm{A},K}\right)^{\mathrm{T}}$, are defined by the relationship

$$f_{\mathrm{A},k} : U_k, R_k \longrightarrow \mathrm{MOS}_k = f_{\mathrm{A},k}(R_k). \tag{10}$$

That means for each application k there is a corresponding application model $f_{\mathrm{A},k}$ available at the optimizer. The application model establishes a relationship between the data rate R_k and a utility metric. As common utility metric the mean opinion score MOS_k is used, defined by the vector

$$\mathbf{MOS} \triangleq \left(\mathrm{MOS}_1, \ldots, \mathrm{MOS}_K\right)^{\mathrm{T}} \tag{11}$$

containing the MOS of all users, which according to Figure 3 is delivered to the optimizer.

The optimizer uses a utility function

$$f_{\mathrm{O}} : f_{\mathrm{A},1}, \ldots, f_{\mathrm{A},K} \longrightarrow f_{\mathrm{O}}\left(f_{\mathrm{A},1}, \ldots, f_{\mathrm{A},K}\right) \tag{12}$$

providing a relationship between applications. The utility function should be symmetric regarding a permutation of its arguments and monotonic for each argument. We decide to maximize the sum of the MOSs of all applications and choose the utility function

$$f_{\mathrm{O}}\left(f_{\mathrm{A},1}, \ldots, f_{\mathrm{A},K}\right) = \sum_{k \in \mathcal{K}} f_{\mathrm{A},k}. \tag{13}$$

Using this utility function, the optimization problem

$$\arg\max_{\{\alpha_1, \ldots, \alpha_K\}} f_{\mathrm{O}}\left(f_{\mathrm{A},1}\left(f_{\mathrm{L},1}(\alpha_1)\right), \ldots, f_{\mathrm{A},K}\left(f_{\mathrm{L},K}(\alpha_K)\right)\right) \tag{14a}$$

subject to

$$0 \leq \alpha_k \ \forall k \in \mathcal{K}, \quad \sum_{k \in \mathcal{K}} \alpha_k = 1 \tag{14b}$$

is to be solved, which delivers $\boldsymbol{\alpha}_{\mathrm{opt}}$ and via (1) also $\mathbf{R}_{\mathrm{opt}}$. The optimizer outputs the resource assignments $\boldsymbol{\alpha}_{\mathrm{opt}}$ and rate allocation $\mathbf{R}_{\mathrm{opt}}$ to the MAC and APP layer, respectively.

3.2.2. Required Overhead. Reviewing the exchanged parameters, we notice that the vectors $\mathbf{R_{max}}$ and $\boldsymbol{\alpha}$ contain only long-term information. No instantaneous CSIT, power allocation, modulation, or schedules have to be exchanged between PHY/MAC layer and the optimizer. Likewise the APP layer model specified in Section 3.1.2 is determined by only two parameters that are slowly time varying. This has the advantage that the system is less sensitive against delays caused by parameter exchange between layers and the optimizer. Robustness against delays is of importance for CLO as base station and application server are most likely located at different physical locations so that control information is to be exchanged over the core network.

If the principles of conventional CLO systems [21] are applied to our case, all considered schedules have to be

Table 1: Number of exchanged parameters.

Number of slots	N_{slot}	52	8
Number of users	K	2	8
Exchanged parameters for:			
all possible schedules	$K^{N_{\text{slot}}+1}+1$	9.0e15	1.3e8
only schedules with different data rates	$K\dfrac{(K+N_{\text{slot}}-1)!}{(N_{\text{slot}})!(K-1)!}+1$	107	5.1e4
model-based proposal	$2K-1$	3	15

transmitted from the link layer to the optimizer. For each schedule at least the K data rates that the users achieve are transmitted. For N_{slot} slots there are

$$K^{N_{\text{slot}}} \tag{15}$$

permutations (each representing one possible schedule). However, since a PGPS scheduler does not utilize channel knowledge, all slots may be considered equally. The scheduler's task is to assign K users to N_{slots} slots (which means to find all combinations of K elements, N_{slots} at a time) whereas one user may be scheduled in multiple slots (repetitions are allowed). Hence, the actual number of schedules is smaller than (15) and is given by [27]

$$\binom{K+N_{\text{slot}}-1}{N_{\text{slot}}} = \frac{(K+N_{\text{slot}}-1)!}{(N_{\text{slot}})!(K-1)!}. \tag{16}$$

This means that for the conventional system [21]

$$K\frac{(K+N_{\text{slot}}-1)!}{(N_{\text{slot}})!(K-1)!} \tag{17}$$

data rate values have to be transmitted to the optimizer and one value is fed back as the chosen schedule.

Table 1 shows some numerical examples for the number of exchanged parameters. Although conventional CLO attains a significant reduction of exchanged parameters by intralayer optimization, which allows to consider only a subset of schedules (16), the control information overhead may still be prohibitive for a high number of users and slots. In contrast, the proposed parameter abstraction needs to transmit only K data rates from the link layer to the optimizer, while $K-1$ values are fed back. Of particular advantage is the fact that the control information overhead is independent of the number of slots N_{slot}.

4. Optimum Resource Assignment

Based on the model-based CLO framework the optimum resource allocation assuming an idealized utility is derived in closed form in this section. The mathematical analysis is the basis of an optimization algorithm presented in Section 5, which maximizes a more realistic utility.

4.1. Problem Statement. The objective is to maximize the sum MOS of all users. With the specific link model (1) and application model (4) the optimization problem (14) can be formulated as follows:

$$\begin{aligned}\boldsymbol{\alpha}_{\text{opt}} &= \arg\max_{\boldsymbol{\alpha}} \sum_{k\in\mathcal{K}} \text{MOS}_k(R_k) \\ &= \arg\max_{\boldsymbol{\alpha}} \sum_{k\in\mathcal{K}} \text{MOS}_k(\alpha_k R_{\max,k})\end{aligned} \tag{18a}$$

subject to

$$0 \le \alpha_k \ \forall k \in \mathcal{K}, \quad \sum_{k\in\mathcal{K}} \alpha_k = 1. \tag{18b}$$

As the above optimization problem is neither convex nor concave, we first define an idealized utility that produces a concave optimization problem.

4.2. Unbounded Application Characteristic. Removing the bounds in the application model (4) results in an unbounded logarithmic relation between utility metric and data rate. The unbounded optimization problem is formulated as:

$$\boldsymbol{\alpha}'_{\text{opt}} = \arg\max_{\boldsymbol{\alpha}} \sum_{k\in\mathcal{K}} \text{MOS}_{0,k} \log \frac{\alpha_k R_{\max,k}}{R_{0,k}} \tag{19a}$$

subject to

$$0 \le \alpha_k \quad \forall\, k \in \mathcal{K}, \tag{19b}$$

$$\sum_{k\in\mathcal{K}} \alpha_k = 1. \tag{19c}$$

The optimization (19a) can be simplified as:

$$\begin{aligned}\boldsymbol{\alpha}'_{\text{opt}} &= \arg\max_{\boldsymbol{\alpha}} \sum_{k\in\mathcal{K}} \text{MOS}_{0,k} \log \alpha_k \\ &= \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \mathbf{MOS}_0)\end{aligned} \tag{20}$$

with the equivalent utility function

$$f(\boldsymbol{\alpha}, \mathbf{MOS}_0) \triangleq \sum_{k\in\mathcal{K}} \text{MOS}_{0,k} \log \alpha_k. \tag{21}$$

The vector $\mathbf{MOS}_0 \triangleq (\text{MOS}_{0,1}, \ldots, \text{MOS}_{0,K})^{\text{T}}$ contains coefficients that characterize the K applications as defined in (4b).

Note that $f(\boldsymbol{\alpha}, \mathbf{MOS}_0)$ and, hence, the solution of the unbounded optimization problem is independent on the physical radio channel, characterized by $R_{\max,k}$, and only depends on $\mathbf{MOS}_0$, which is determined by the ratio between $R_{1.0,k}$ and $R_{4.5,k}$.

For finding a closed form solution of the optimum resource assignment $\boldsymbol{\alpha}'_{\text{opt}}$ in (19), in the following we prove the concavity of the optimization problem, derive the optimum share of resources between two users, and find a solution for the absolute resource share of a user.

Reformulating the constraint (19c) as:

$$\alpha_\ell = 1 - \sum_{\substack{k\in\mathcal{K} \\ k\neq\ell}} \alpha_k, \quad \ell \in \mathcal{K} \tag{22}$$

and inserting the result into (21) yields

$$f(\boldsymbol{\alpha}, \mathbf{MOS}_0) = \sum_{\substack{k\in\mathcal{K}\\ k\neq\ell}} \mathrm{MOS}_{0,k} \log \alpha_k + \mathrm{MOS}_{0,\ell} \log\left(1 - \sum_{\substack{n\in\mathcal{K}\\ n\neq\ell}} \alpha_n\right). \tag{23}$$

Now, the first and second partial derivatives in directions of α_k and α_m can be determined,

$$\left.\frac{\partial f}{\partial \alpha_k}\right|_{k\neq\ell} = \frac{\mathrm{MOS}_{0,k}}{\alpha_k} - \frac{\mathrm{MOS}_{0,\ell}}{1 - \sum_{n\in\mathcal{K}, n\neq\ell} \alpha_n}, \tag{24}$$

$$\left.\frac{\partial^2 f}{\partial \alpha_k^2}\right|_{k\neq\ell} = -\frac{\mathrm{MOS}_{0,k}}{\alpha_k^2} - \frac{\mathrm{MOS}_{0,\ell}}{\left(1 - \sum_{n\in\mathcal{K}, n\neq\ell} \alpha_n\right)^2}, \tag{25}$$

$$\left.\frac{\partial^2 f}{\partial \alpha_k \partial \alpha_m}\right|_{k,m\neq\ell, k\neq m} = -\frac{\mathrm{MOS}_{0,\ell}}{\left(1 - \sum_{n\in\mathcal{K}, n\neq\ell} \alpha_n\right)^2}. \tag{26}$$

Considering (4b) and (4d), it follows that $\mathrm{MOS}_{0,k} > 0 \ \forall k \in \mathcal{K}$ so that

$$\frac{\partial^2 f}{\partial \alpha_k^2} < 0 \quad \forall k \in \mathcal{K} \tag{27}$$

and

$$\frac{\partial^2 f}{\partial \alpha_k \partial \alpha_m} < 0 \quad \forall k \in \mathcal{K}. \tag{28}$$

This means that the graph is strictly concave downwards and any extremum not being located on the domain borders maximizes the utility. Therefore, provided for all $k \in \mathcal{K}$ the following condition is satisfied

$$\left.\frac{\partial f}{\partial \alpha_k}\right|_{k\neq\ell} = 0, \tag{29}$$

the global maximum is found. Setting (24) to zero yields

$$\left(\frac{\mathrm{MOS}_{0,\ell}}{\mathrm{MOS}_{0,k}} + 1\right)\alpha_k = 1 - \sum_{\substack{n\in\mathcal{K}\\ n\neq\ell,k}} \alpha_n. \tag{30}$$

Likewise, the optimum share for user ℓ, α_ℓ, when α_k is fixed, is determined by differentiating (24) with respect to α_ℓ and setting the result to zero, which corresponds to swapping users k and ℓ in (30). By combining the result with (30) the dependency to other users $n \neq k, \ell$ disappears. This means that the relation between the optimum resource assignments of any two users, k and ℓ, is independent of all other users' utility functions. After some algebraic manipulations the relation

$$\alpha_k = \frac{\mathrm{MOS}_{0,k}}{\mathrm{MOS}_{0,\ell}} \alpha_\ell \tag{31}$$

between the optimization coefficients of users k and ℓ is obtained.

For finding an absolute value for the optimization coefficients $\boldsymbol{\alpha}$ the relation (31) is inserted into the constraint (19c), which yields

$$\alpha_\ell = \frac{\mathrm{MOS}_{0,\ell}}{\sum_{k\in\mathcal{K}} \mathrm{MOS}_{0,k}} \tag{32}$$

as the final solution of the unbounded optimization problem (19).

As a special case it can be easily seen from (32) that if all users have the same parameter MOS_k, then the resources are distributed equally to the users,

$$\mathrm{MOS}_{0,1} = \cdots = \mathrm{MOS}_{0,K} \Longrightarrow \alpha_k = \frac{1}{K} \quad \forall k \in \mathcal{K}. \tag{33}$$

Interestingly, given that all users use the same application, the optimum resource allocation for the unbounded problem results in an equal resource scheduler where all users are assigned the same number of slots. This implies that users experiencing a good channel receive higher data rates and therefore enjoy better QoS, as adaptive transmission is more bandwidth efficient in this case.

In summary, the optimum resource allocation for the unbounded optimization problem (32) is independent of the channel conditions; the number of assigned slots (the allocated bandwidth) is exclusively determined by the application characteristics; users with a good channel enjoy higher data rates. On the other hand, all users are given a fair share of the available resources. This is in a sharp contrast to a maximum throughput scheduler, which exclusively serves good users while users experiencing a poor channel starve for resources. The significance of this finding is that the maximized utility in (19) is an idealized measure of user-perceived quality.

4.3. Subset of Users. For solving the bounded optimization problem (18), it is useful to solve the unbounded problem only for a subset of "variable" users $\mathcal{K}_{\mathrm{var}} \in \mathcal{K}$. The remaining users $\mathcal{K}_{\mathrm{fix}} = \mathcal{K} \setminus \mathcal{K}_{\mathrm{var}}$ have fixed optimization coefficients α_k and are not subject to optimization. Here, the notation $\mathcal{K} \setminus \mathcal{K}_{\mathrm{var}}$ denotes the relative complement of set $\mathcal{K}_{\mathrm{var}}$ in set $\mathcal{K}$.

The constraint (19c) is rewritten as

$$\sum_{k\in\mathcal{K}_{\mathrm{var}}} \alpha_k = 1 - \sum_{m\in\mathcal{K}_{\mathrm{fix}}} \alpha_m. \tag{34}$$

Following the derivation in Section 4.2, inserting (31) gives

$$\sum_{k\in\mathcal{K}_{\mathrm{var}}} \frac{\mathrm{MOS}_{0,k}}{\mathrm{MOS}_{0,\ell}} \alpha_\ell = 1 - \sum_{m\in\mathcal{K}_{\mathrm{fix}}} \alpha_m, \tag{35}$$

which finally yields

$$\alpha_\ell = \left(1 - \sum_{m\in\mathcal{K}_{\mathrm{fix}}} \alpha_m\right) \frac{\mathrm{MOS}_{0,\ell}}{\sum_{k\in\mathcal{K}_{\mathrm{var}}} \mathrm{MOS}_{0,k}}. \tag{36}$$

5. Optimization Algorithm Maximizing the User-Perceived Quality

Based on the analytical solution for the unbounded problem in Section 4, an optimization algorithm for the bounded problem (18) is presented in this section. In an intermediate step a solution for the upper bounded problem is derived, where the application characteristic $\mathrm{MOS}_k(R_k)$ is upper bounded at an MOS of 4.5. Then the solution of the bounded problem is developed, and its computational complexity is assessed. Finally, the proposed CLO algorithm is extended to support different priority classes.

5.1. Upper Bounded Problem. We define the upper bounded application characteristic by

$$\mathrm{MOS}_k^{\mathrm{u}}(R_k) = \begin{cases} \mathrm{MOS}_{0,k} \log \dfrac{R_k}{R_{0,k}} & : R_k < R_{4.5,k}, \\ 4.5 & : R_k \geq R_{4.5,k}, \end{cases} \tag{37}$$

which gives the upper bounded optimization problem

$$\arg\max_{\boldsymbol{\alpha}} \sum_{k \in \mathcal{K}} \mathrm{MOS}_k^{\mathrm{u}}(\alpha_k R_{\max,k}) \tag{38a}$$

subject to

$$0 \leq \alpha_k \ \forall k \in \mathcal{K}, \quad \sum_{k \in \mathcal{K}} \alpha_k = 1. \tag{38b}$$

Let $R'_{\mathrm{opt},k} = \alpha'_{\mathrm{opt},k} R_{\max,k}$ denote the optimum rate allocation of user k of the unbounded problem (32). In case $R'_{\mathrm{opt},k} > R_{4.5,k}$, the rate for user k may be reduced to $R_{4.5,k}$ without sacrificing service quality, and the retained resources can be given to users with $R'_{\mathrm{opt},\ell} < R_{4.5,\ell}$, $\ell \neq k$. A solution of this concave problem is found by the iterative algorithm:

Step 1. Initially, $\mathcal{K}_{\mathrm{fix}} = \varnothing$ and $\mathcal{K}_{\mathrm{var}} = \mathcal{K}$.

Step 2. Solve unbounded problem (36).

Step 3. Users with $R'_{\mathrm{opt},k} \geq R_{4.5,k}$ are moved from $\mathcal{K}_{\mathrm{var}}$ to $\mathcal{K}_{\mathrm{fix}}$ and set $\alpha_k = R_{4.5,k}/R_{\max,k}$.

Step 4. If any user has been moved in Step 3, continue with Step 2, otherwise stop.

If any of the application characteristics deviates from (4), Step 2 can be replaced by a conventional algorithm that solves the unbounded problem. Alternatively, appropriate values for $R_{1.0,k}$ and $R_{4.5,k}$ can be chosen to approximate the real application characteristic, giving rise to a certain deviation to the exact solution. Optionally, this approximation could be used as a starting point for an applicable conventional algorithm.

5.2. Bounded Problem. We approach the bounded optimization problem (18) by dividing it into two subproblems: first, a subset of users is determined who cannot be served and therefore get no resources, $\alpha_k = 0$; second, for the remaining users the upper bounded optimization problem from Section 5.1 is solved. In case dropped users are selected appropriately in the first step, the remaining served users will always achieve data rates $R_k > R_{1.0,k}$ so that the solution for the bounded problem is optimum.

The following iterative algorithm for the solution of the bounded problem is formulated as follows.

Step 1. Initially, all users are served.

Step 2. Drop users as detailed in Steps 2.1–2.4.

Step 2.1. If `stop_criterion` is fulfilled, continue with Step 3.

Step 2.2. Solve upper bounded problem for the served users as described in Section 5.1.

Step 2.3. User $k_{\mathrm{drop}} = \arg\max_k \sum_{k' \neq k} \mathrm{MOS}_{k'}^{\mathrm{u}}$ is dropped by setting $\alpha_{k_{\mathrm{drop}}} = 0$.

Step 2.4. Continue with Step 2.1.

Step 3. Solve upper bounded problem for the served users as described in Section 5.1 and stop.

In this algorithm the `stop_criterion` determines how many users are served. When the objective is to maximize the sum of all users' MOS, referred to as "increase sum MOS", an appropriate strategy is to continue dropping users until this does not further improve the sum MOS.

An alternative `stop_criterion` is to check

$$\alpha_k - \alpha_{\mathrm{stop},k} > 0 \quad \forall k, \tag{39a}$$

where

$$\alpha_{\mathrm{stop},k} \triangleq (\alpha_k \mid \mathrm{MOS}_k(\alpha_k) = \mathrm{MOS}_{\mathrm{stop},k}). \tag{39b}$$

This condition checks whether the MOS that would be achieved with the allocated resources α_k exceeds a certain minimum $\mathrm{MOS}_{\mathrm{stop},k} \in [1, 4.5]$. Setting $\mathrm{MOS}_{\mathrm{stop},k} = 1 \ \forall k \in \mathcal{K}$ ensures that only a minimum of users are dropped, while no resources are wasted to users that would anyhow experience unacceptable service quality of $\mathrm{MOS}_k(\alpha_k) = 1$. On the other hand, higher values of $\mathrm{MOS}_{\mathrm{stop},k}$ enforce a certain minimum perceived quality. This variant of the algorithm is therefore termed "reduce outage".

As the above discussion touches upon the issue of admission control, other criteria that determine which users are admitted to the system might be introduced. For example, in a cellular system it might be desirable to give priority to users that hand over from a neighboring cell rather than to serve a user who wishes to enter the network.

5.3. Computational Complexity. An appealing feature is that the proposed optimization algorithm deterministically terminates after a certain time. To prove this the worst case run time is calculated in the following. Since in each iteration at least one user is dropped, there are at most K iterations

in the outer loop. The inner loop computes the solution of the upper bounded problem. In the worst case, one user is moved from $\mathcal{K}_{\text{var}}$ to $\mathcal{K}_{\text{fix}}$ so that the number of iterations at most equals the number of served users. The total number of iterations is therefore upper bounded by $K(1+K)/2$.

An observation from the simulation results in Section 6 is that typically most users can transmit. Hence, the number of iterations for the outer loop is likely to be significantly smaller than K. Likewise, trials suggest that for the inner loop it is rather unlikely that more than two iterations are required. Since the essential calculation within the inner loop is given by the closed form expression (36), the total complexity of the optimization algorithm is low.

5.4. Priority Classes. In order to support different priority classes, the utility function is adjusted in the following.

Let $\lambda_k \in \mathbb{R}$ be a real number that reflects the priority of user k where, without loss of generality, $\lambda_k > \lambda_\ell$ indicates that user k has a higher priority than user ℓ. Priority classes are incorporated to the utility function by substituting the application dependent constant $\text{MOS}_{0,k}$ in (19) by the function $g_k(\text{MOS}_{0,k}, \lambda_k)$, that is,

$$\sum_{k\in\mathcal{K}} g_k(\text{MOS}_{0,k}, \lambda_k) \log \frac{\alpha_k R_{\max,k}}{R_{0,k}}. \tag{40}$$

In the calculation of the first and second partial derivatives in direction of α_k and α_m in (24), (25), and (26), $\text{MOS}_{0,k}$ is treated as a constant. Therefore, the derivation of the unbounded optimization problem in Section 4.2 also applies to the priority function $g_k(\text{MOS}_{0,k}, \lambda_k)$, if the following condition holds

$$\frac{\partial g_k(\text{MOS}_{0,k}, \lambda_k)}{\partial \alpha_\ell} = 0 \quad \forall \{k, \ell\} \in \mathcal{K}^2. \tag{41}$$

Likewise, (4b) and (4d) strictly require a positive constant $\text{MOS}_{0,k}$, which translates to

$$g_k(\text{MOS}_{0,k}, \lambda_k) > 0 \quad \forall k \in \mathcal{K}. \tag{42}$$

Under these conditions, the conclusions from Section 4.2 apply: the utility function that supports priority classes (40) is strictly concave downwards, and the underlying optimization problem is solved by substituting $\text{MOS}_{0,k}$ with $g_k(\text{MOS}_{0,k}, \lambda_k)$ in (31), (32), and (36).

An intuitive realization of a priority function that satisfies the constraints (41) and (42) is given by

$$g_k(\text{MOS}_{0,k}, \lambda_k) = \lambda_k \text{MOS}_{0,k}, \quad \lambda_k > 0 \ \forall k \in \mathcal{K}, \tag{43}$$

which is similar to the approach described in [19]. This function is applied for obtaining the numerical results presented in Section 6.5.

There are several possibilities how to further incorporate priority classes, for example, by adjusting the upper bound of the upper bounded optimization problem, the stop criterion or by using an alternative criterion for dropping users.

Table 2: Link layer parameters.

Transmission scheme	OFDMA
Number of subcarriers	$N = 416$
Cyclic prefix duration	$3.2\,\mu$s
Symbol mapping	BPSK, 4-, 16-, 64-QAM
Channel coding	Convol., $R_c \in \left\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{9}{16}, \frac{2}{3}, \frac{3}{4}\right\}$
Channel bandwidth	$B = 16.25$ MHz
Channel model	WINNER urban macro-cell [28]
Duplex ratio DL/UL	1/1
Cell radius	$50 \cdots 500$ m
Shadowing	log-normal, $\sigma_s = 8$ dB
Path loss	$38.4\,\text{dB} + 35.0\,\text{dB}\log_{10}(d/\text{m})$
Center frequency	$f_0 = 5.25$ GHz
Transmit power	10 W
Antenna gain	8 dBi
Noise figure	7 dB
Noise spectrum density	-174 dBm/Hz
Delay spread	$\tau_{ds} = 313$ ns
Maximum Doppler speed	$v = 50$ km/h
Slot size (freq. $\times$ time)	8×12
Number of users	$K = 1, \ldots, 64$
Number of available slots	$N_{\text{slot}} = 52$
Scheduler	PGPS

6. Performance Evaluation

The performance of the proposed CLO framework is evaluated by means of system simulations. The link layer parameters listed in Table 2 mostly follow the WINNER (World Wireless Initiative New Radio, URL: www.ist-winner.org) system concept [2].

6.1. Simulation Setup. We consider an OFDMA downlink that occupies a bandwidth of $B = 16.25$ MHz. Due to the inherent orthogonality of orthogonal frequency division multiplexing (OFDM), each subcarrier in each OFDM symbol may be assigned to a different user without causing interference, so that users can be scheduled independently in time and frequency. Adjacent subcarriers and OFDM symbols are correlated and, therefore, experience a similar channel gain. In order to limit the signaling overhead 8×12 symbols are grouped to form one slot.

The WINNER typical urban macrocell channel (model C2 [28]) is used, which models channel attenuation due to frequency selective fading, distance dependent path loss and log-normal shadowing [29]. Instantaneous channel variations due to velocities of mobile users are generated using Jakes' model [30]. The channel model is implemented such that the average SNR always allows transmission with the lowest supported modulation and coding scheme. This is motivated by the fact that users with lower SNR would not be able to establish a connection to the base station and, hence, cannot request to be served. While the average SNR

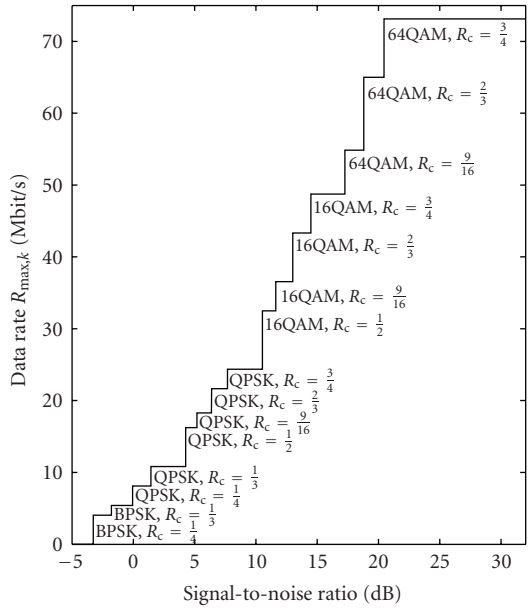

Figure 6: Adaptive modulation: relation between instantaneous data rate and signal-to-noise ratio (SNR).

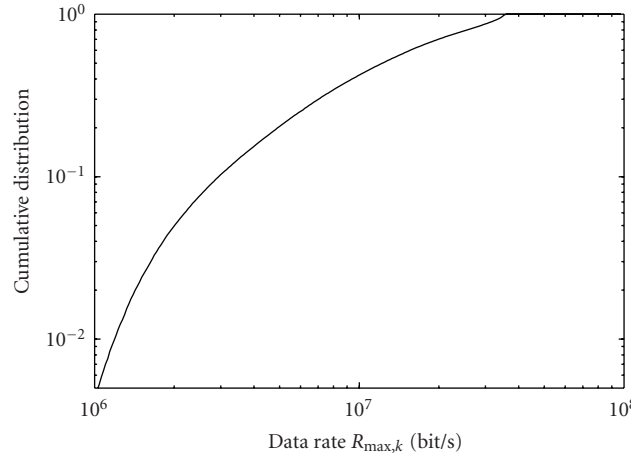

Figure 7: CDF of maximum data rate $R_{\max,k}$, which characterizes the communications channel on the link layer.

always exceeds the given limit, the instantaneous SNR may be significantly lower due to frequency selective fading.

Mobile velocities up to $v = 50$ km/h are assumed, which implies that instantaneous CSIT may not be available. It is assumed that the average SNR over all simultaneously transmitted slots is available for link adaptation. Hence, the same modulation and coding scheme is applied to all subcarriers of one user during one slot duration. However, slots assigned to different users will typically use a different modulation and coding scheme.

The transmitter chooses the symbol mapping with cardinality M and code rate R_c of a convolutional code, based on the average SNR of each user k (see Figure 6). Note that due to half-duplex transmission the average data rate is only half of the instantaneous data rates indicated in Figure 6. The modulation and coding scheme is selected that achieves the largest spectral efficiency $\eta_k = R_c \log_2 M$ at a frame error rate (FER) of 10^{-2}. The SNR values for which FER $= 10^{-2}$ are determined by reference simulations and are stored in a look-up table. It is assumed that an ARQ protocol at the link layer takes care of error events by retransmitting erroneously received packets. Due to the low occurrence of errors at FER $= 10^{-2}$ retransmissions only have marginal impact on the throughput and will therefore not affect the perceived quality. Hence, simulations assume that packets are always received error free.

For CLO the long-term average data rate $R_{\max,k} = \eta_k N_{\text{slot}}$ for each user k indicates the link capacity and is the relevant abstraction of the link layer. Figure 7 shows the cumulative distribution function (CDF) of $R_{\max,k}$, which is averaged over a large number of randomly chosen channel realizations and user locations within a cell.

Simulations are executed as follows: every 100 milliseconds independent shapshots of path loss and shadowing realizations are generated for each user according to a uniform user distribution within the cell area. Then $\mathbf{R}_{\max}$ is estimated and passed to the optimizer. CLO is performed to determine the optimum share of resources $\boldsymbol{\alpha}_{\text{opt}}$, which is subsequently fed back to the PGPS scheduler at the MAC layer.

After the 100-millisecond snapshot, the actually achieved average data rates are determined. The actually achieved data rates may deviate from the optimizer's estimate $\mathbf{R}_{\max}$. Each user's MOS is determined based on the user's application and the achieved data rate. Then, the CDF of the MOS averaged over all users is calculated.

6.2. Performance of Different Optimization Algorithms. In Figure 8, the CDF of the MOS is shown for the different resource allocation strategies and optimizer variants discussed in Section 5. The applications of all $K = 16$ users are described by the same parameters $R_{1.0} = 100$ kbit/s and $R_{4.5} = 1$ Mbit/s (compare Figure 5). As a reference equal resource allocation with $\alpha_k = 1/16$ for all 16 users is also plotted, which is the optimum resource assignment of the unbounded optimization problem (19) (see Section 4.2). Greedy resource allocation [19], as a conventional technique for solving optimization problems, is also included for comparison. From our experience the Greedy algorithm is significantly more computationally expensive than the proposed CLO algorithm. The other two curves show the performance of the proposed algorithm, the "increase sum MOS," and the "reduce outage" variants, where the stop criterion is set to $\text{MOS}_{\text{stop},k} = 1 \ \forall k \in \mathcal{K}$.

As seen in Figure 8, both variants outperform equal resource allocation and achieve a comparable average MOS as greedy resource allocation. Compared to equal resource allocation, any performance improvement of the considered optimization algorithms is due to the bounds in the MOS trajectory, since users with $R_k = R_{\max,k}/16 > R_{4.5}$ perceive the same QoS as if they were served with the reduced rate $R'_k = R_{4.5}$. Likewise, users with $R_k < R_{1.0}$ perceive the same QoS as a user who is not served at all. The "reduce outage"

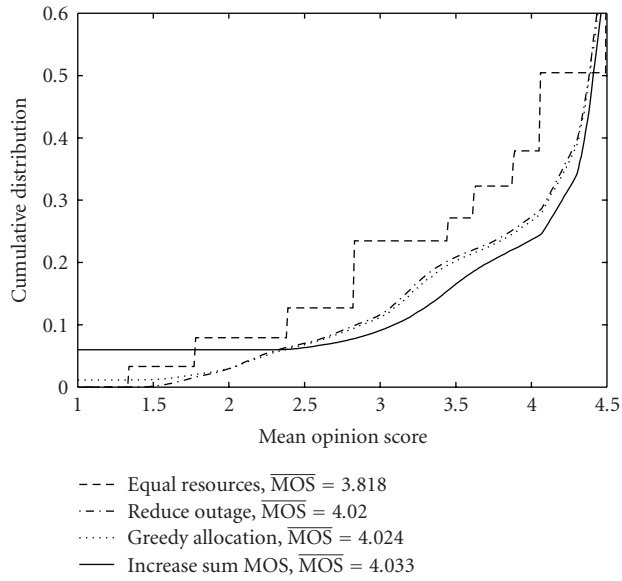

Figure 8: CDF of perceived quality for different optimization algorithms.

variant serves practically all users, although some perceive a poor service quality. In contrast, the "increase sum MOS" variant tends to drop users with poor quality and assigns the freed resources to served users. This is due the objective, which aims to maximize the sum MOS of all users: a user will be dropped, if the increase in MOS of the served users outweighs the decrease in MOS of dropping a certain user.

6.3. Deviation due to Application Model Abstraction. In Section 6.2, the application is characterized by the idealized bounded logarithmic relationship (4), so that the APP layer model at the optimizer perfectly matches the application characteristics. In order to assess the benefits of CLO in a real system with real applications running, a video streaming example is chosen where the user-perceived quality is approximated as described in Section 2.2.1. Eight different H.264/MPEG-4 AVC videos in common intermediate format (CIF) resolution at 30 Hz frame rate are cut into snippets containing one group of pictures (GOP) each. With a GOP size of 32 frames the snippets contain approximately 1 second of video. Further parameters of the videos are summarized in Tables 3 and 4. The snippets are subsequently analyzed to extract the parameters a, b, and c for each snippet.

In order to assess the effect of rate variations of the video stream over time, for each 100-millisecond PHY channel snapshot a new (random) snippet of the respective video stream is used. For the proposed optimization algorithm from Section 5 the parameters $R_{1.0,k}$ and $R_{4.5,k}$ are estimated by the application server for each video snippet and provided to the CLO. Because the optimization algorithm is based on the bounded logarithmic relationship (4), which deviates from the actually used video model (3), the decided resource distribution will be suboptimum. For comparison CLO with greedy optimization using the exact video model (3) is also simulated.

Table 3: Video parameters.

Video coding	H.264/MPEG-4 AVC [7]
Implementation	JSVM 9.12.2, 25 April 2008 [31]
Resolution	CIF (352 × 288)
Frame rate	30 Hz
Chroma subsampling	4:2:0
GOP size	32
GOP coding structure	I-P-· · · -P
PSNR range	$PSNR_{1.0}$ = 30 dB, $PSNR_{4.5}$ = 42 dB

Table 4: Transmitted video streams.

Video name	Duration (GOP)	Average desired data rate $\overline{R}_{4.5}$	Ratio $\overline{R}_{4.5}/\overline{R}_{1.0}$
Foreman	9	2,156 kbit/s	18
Mother	9	447 kbit/s	26
News	9	638 kbit/s	11
Container	9	1,159 kbit/s	22
Salesman	9	2,265 kbit/s	40
Bus	4	4,141 kbit/s	7
City	9	2,202 kbit/s	13
Crew	9	2,677 kbit/s	15

As seen in Figure 9, for the considered real video streams similar conclusions as for the generic applications from Section 6.2 can be drawn. The considered optimizers exhibit similar performance, achieving a significantly superior MOS with respect to equal resource allocation.

6.4. Guaranteed Service Quality. It may be desirable to support the demand for minimum QoS. This may be accomplished by tuning the parameter MOS_{stop} of the stop criterion in the "reduce outage" variant of the proposed optimization algorithm. As the stop criterion controls which users are dropped from the list of active users (see Section 5.2), setting MOS_{stop} to a value in the range $[1, 4.5]$ ensures that all served users achieve at least a minimum perceived quality of MOS_{stop}.

Figure 10 shows the CDF of the achieved sum MOS for $MOS_{stop} = 2.0$ and $MOS_{stop} = 3.0$. The higher MOS_{stop} the less users achieve the required data rates due to poor channel conditions and are therefore not served. On the other hand, the served users with better channels benefit from freed resources of the dropped users, which improves their perceived quality.

Figure 11 shows the MOS, averaged over all users and channel realizations, against MOS_{stop}. The choice of MOS_{stop} affects the overall perceived quality and the maximum is approached for $MOS_{stop} \approx 2$. In case $MOS_{stop} < 2$, users with poor channels are served, which have only a marginal contribution to the overall sum MOS. On the other hand, if $MOS_{stop} > 2$, an increasing number of users are denied service, which cannot be compensated by the enhanced QoS

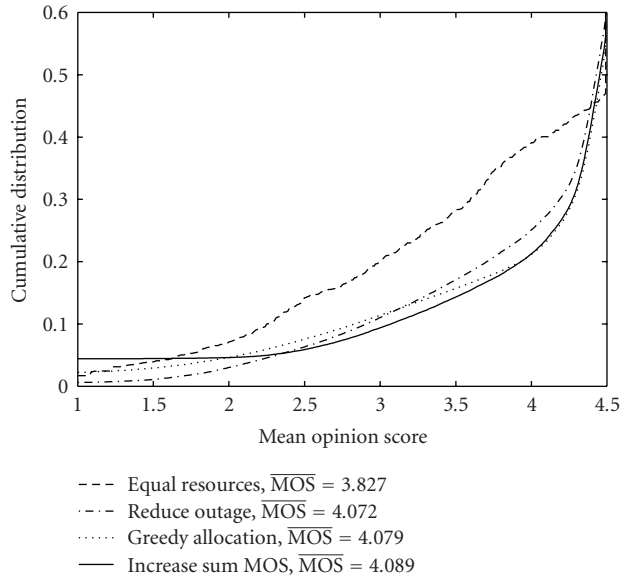

Figure 9: CDF of the perceived quality for video streaming with nonconcave application characteristic.

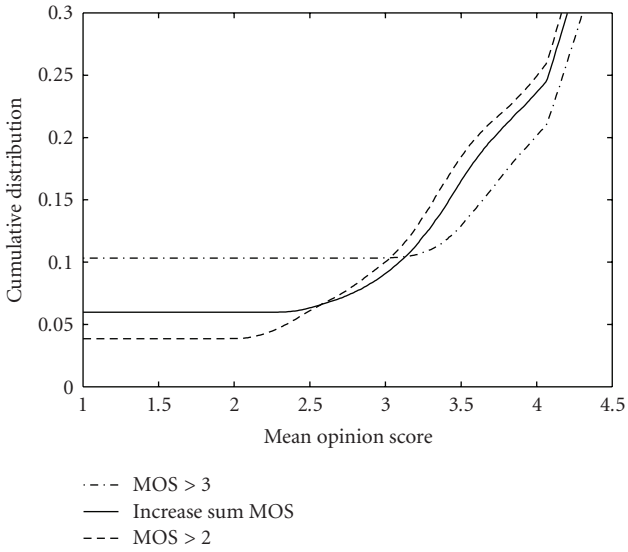

Figure 10: CDF of the perceived quality for different minimum MOS constraints MOS_{stop}. For comparison the "increase sum MOS" variant is also included.

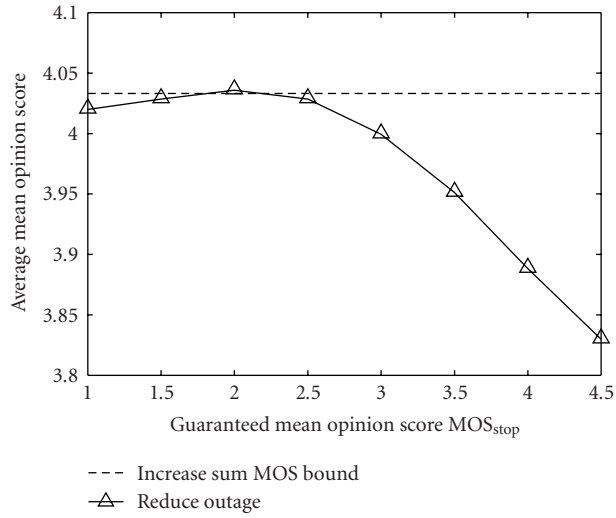

Figure 11: Average MOS as a function of the minimum MOS constraints MOS_{stop}.

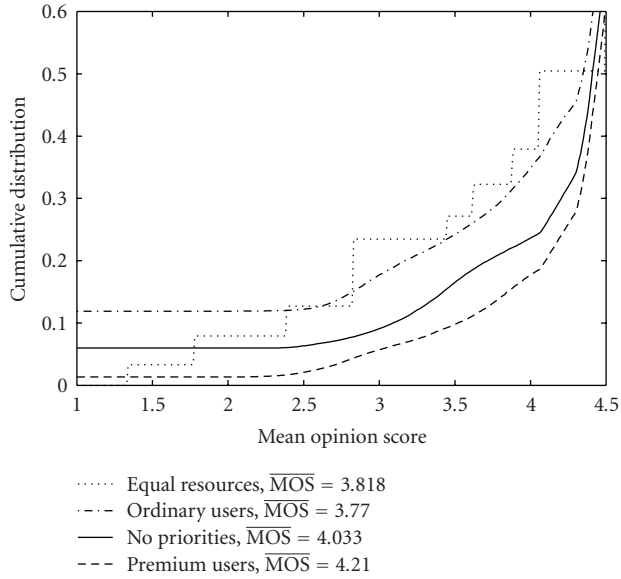

Figure 12: CDF of the perceived quality for ordinary and premium traffic.

of the remaining active users. The perceived quality achieved by the "increase sum MOS" variant, which approximates the maximum sum MOS, is also indicated in Figure 10.

6.5. Traffic Priority Classes. The performance of CLO supporting different traffic priority classes developed in Section 5.4 is examined in Figure 12. The $K = 16$ users, all running the same applications, are split up into two priority groups of 8 users each; premium and ordinary users are given a priority of $\lambda_k = 2$ and $\lambda_k = 1$, respectively.

Figure 12 shows the CDF of the sum MOS. Premium users exhibit a significantly better MOS than ordinary users and are more likely to be served.

6.6. Application Characteristic. In order to identify for which application characteristics CLO is most effective, different generic application classes are examined, characterized by their relationship between data rate and perceived quality as described by the parameters $R_{1.0}$ and $R_{4.5}$ (see Figure 5), for the "increase sum MOS" variant of the proposed optimization algorithm. The application characteristics are the same for all users and $R_{4.5} = 10R_{1.0}$ is chosen. Figure 13 shows the average MOS against the required data rate for a maximum perceived quality of $R_{4.5}$, for a system with $K = 16$ users.

As seen from Figure 13, the attainable gains of CLO maximizing the sum MOS (solid lines) over equal resource

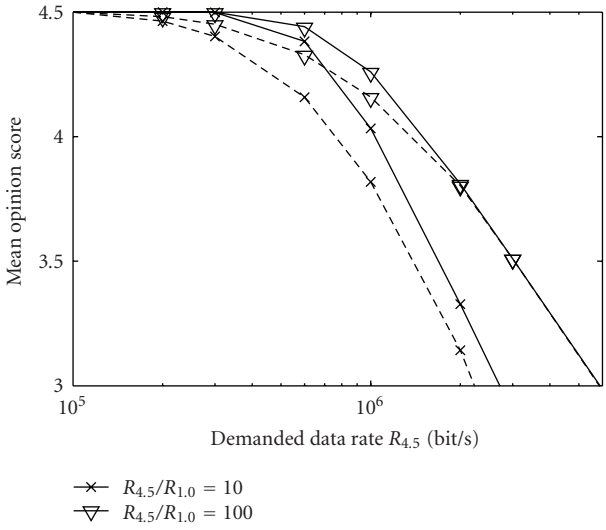

Figure 13: Impact of rate-distortion characteristic on the average MOS. Solid and dashed lines show results for the proposed CLO and equal resource allocation, respectively.

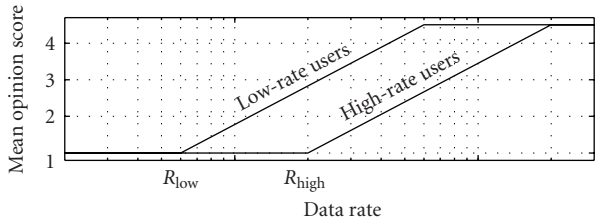

Figure 14: Example characteristic for two user groups running different application classes.

allocation (dashed lines) are dependent on both $R_{4.5}$ and the ratio $R_{4.5}/R_{1.0}$. For low data rate requirements the CLO gain diminishes, as there is an excess of available resources to serve all users with excellent quality MOS = 4.5. For increasing data rate requirements the CLO gain depends on the ratio $R_{4.5}/R_{1.0}$, in the way that the CLO gain increases with decreasing $R_{4.5}/R_{1.0}$. This is explained by the fact that for an increasing ratio $R_{4.5}/R_{1.0}$ the MOS characteristic as a function of the data rate, $\text{MOS}_k(R_k)$ in (4), approaches the unbounded problem addressed in Section 4.2, for which according to (33) equal resource allocation is optimum. In other words, the attainable CLO gains over equal resource allocation with $\alpha_k = 1/K$ are due to users whose rates $R_k = R_{\max,k}/K$ are outside the logarithmic range of $\text{MOS}_k(R_k)$. As the logarithmic range is specified by the ratio $R_{4.5}/R_{1.0}$, the lower $R_{4.5}/R_{1.0}$ the higher the gains to be achieved by optimization.

6.7. Mixed Service Classes. In Figures 14–16 a scenario with two user groups is investigated. Each of the two user groups run applications of a different service class, characterized by different data rate requirements, as illustrated in Figure 14. Low- and high-rate users request a minimum data rate $R_{1.0} = R_{\text{low}}$ and $R_{1.0} = R_{\text{high}}$, respectively.

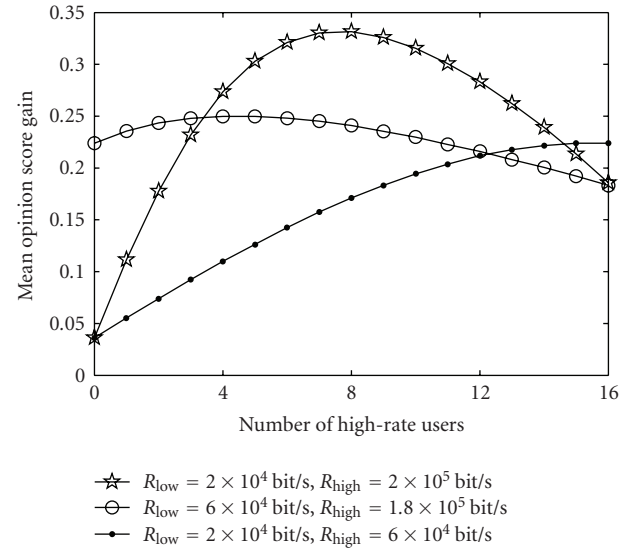

Figure 15: Gain in average MOS for different ratios of low-rate and high-rate users.

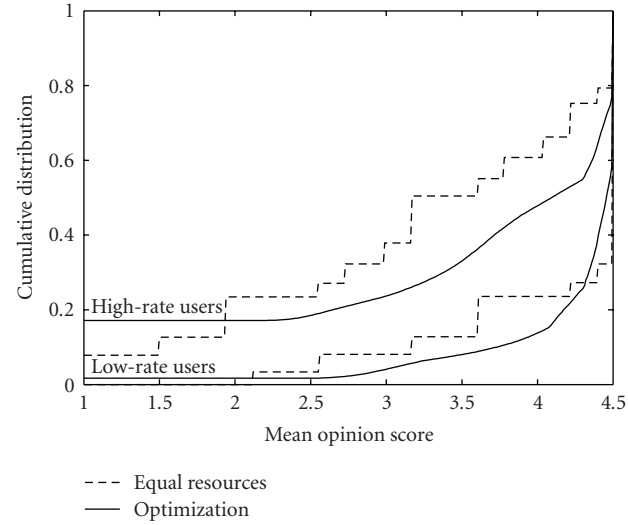

Figure 16: CDF of the perceived quality for two application classes. Both high- and low-rate users benefit from CLO.

Figure 15 shows the CLO gain in sum MOS relative to equal resource allocation against the number of users in each group. Results are plotted for different values of R_{high} and R_{low}, for a total number of $K = 16$ users and $R_{4.5}/R_{1.0} = 10$. Interestingly, in some cases the overall MOS gain for scenarios with mixed service classes exceeds the case when all users are within either of the service classes. This is due to the freed resources by replacing a high-rate user by a less demanding low-rate user, which allows the remaining users to fetch some of the freed resources.

The relationship between low- and high-rate users is further investigated in Figure 16, which shows the CDF of the sum MOS for both user groups. Corresponding to the

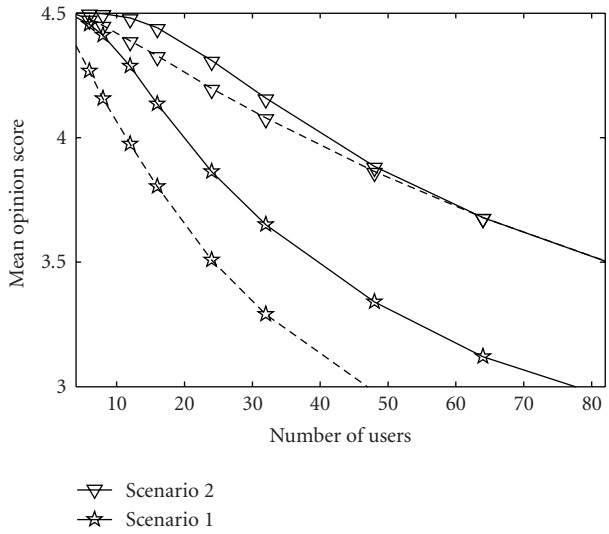

Figure 17: Overall gain achieved by CLO in terms of number of served users K. Solid and dashed lines correspond to the CLO variant "increase sum MOS" and equal resource allocation, respectively.

maximum in Figure 15, there are 4 users with $R_{\text{low}} = 6 \times 10^4$ bit/s and 12 users with $R_{\text{high}} = 1.8 \times 10^5$ bit/s. An appealing observation is that both user groups gain from CLO. While the average gain is $\Delta\text{MOS} = 0.22$, low- and high-rate users gain $\Delta\text{MOS} = 0.10$ and $\Delta\text{MOS} = 0.26$ in overall perceived quality, respectively.

6.8. System Performance. In order to assess the attainable MOS gains from a system level perspective, the average sum MOS is plotted against the number of users K in Figure 17. Two scenarios are investigated. In scenario 1, there are two groups with equal number of users, where low- and high-rate users request the rate $R_{\text{low}} = 2 \times 10^4$ bit/s and $R_{\text{high}} = 2 \times 10^5$ bit/s, respectively. It can be deduced from Figure 17 that CLO maximizing the sum MOS (solid lines) increases the number of users being served with the same average perceived quality by more than 60%, compared to equal resource allocation (dashed lines). In scenario 2, all users run the same application with a desired data rate of $R_{4.5} = 6 \times 10^5$ bit/s and $R_{4.5}/R_{1.0} = 100$, which achieves a comparably small MOS gain of at most $\Delta\text{MOS} = 0.11$, as reported in Section 6.6 for $K = 16$ users. In scenario 2, CLO also enables to serve more users with the same perceived quality, although in this case the gains diminish for increasing number of users. In line with the discussion in Section 6.6, for scenario 2 gains of CLO over equal resource allocation are mainly in the region where the sum MOS is high, since then users, whose rate $R_k = R_{\max,k}/K$ is outside the logarithmic range of $\text{MOS}_k(R_k)$ in (4), are more likely. Otherwise, equal resource allocation tends to approach the optimum resource allocation strategy, leading to diminishing CLO gains.

7. Conclusion

Resource allocation with QoS constraints where multiple users share a wireless downlink is one key challenge in the design of future wireless systems. The MOS is chosen as a universal utility metric for the user-perceived quality for CLO between link and APP layer.

Adaptive transmission based on long-term CSIT over a time and frequency selective fading channel is considered, including distance dependent path loss and log-normal shadowing. Applications are described by a rate-distortion characteristic, expressed by the MOS. With these settings a model-based CLO framework is devised, which emulates the functionalities of the system layers within the optimizer. Compared to known CLO approaches significantly less parameters need to be exchanged. Simulations of a video streaming scenario confirm that model mismatch, where the APP layer model at the optimizer is not perfectly matched to the actual application, only results in modest performance degradation.

As a metric for the user satisfaction we chose to maximize the sum MOS, which resulted in a nonconcave optimization problem. Given an idealized utility metric with an unbounded logarithmic relation between perceived quality and data rate, a concave problem is retained, so that the optimum resource allocation is derived in closed form. One noteworthy result of the analysis is that the optimum solution is independent of the physical channels and is solely described by the application characteristics.

The theoretical findings are the basis for a low complexity and easy to implement CLO algorithm for the more realistic nonconcave optimization problem. The proposed iterative optimization algorithm is significantly less complex than known optimization algorithms and has the appealing feature to deterministically terminate.

The proposed algorithm offers an additional degree of freedom to the network operator to configure its own policies, such as enhancing user satisfaction, ensuring a minimum perceived quality to all users, or to operate the wireless system with higher load so as to maximize revenue. Furthermore, different priority classes can be supported.

The attainable gains of CLO strongly depend on the application characteristics. The higher the sensitivity of the perceived quality to changes of the data rate, the more considerable the gains that can be achieved. Dependent on the application more than 60%, additional users can be served without sacrificing user satisfaction. If multiple service classes with different application characteristic are running simultaneously, all users can be expected to benefit from CLO. In some cases additional CLO gains that exploit a certain mix of service classes are observed.

Acknowledgment

This paper was presented in part at the IEEE Int. Conf. on Communications (ICC'2007), Glasgow, UK, at the IEEE Int. Symp. on Wireless Communication Systems (ISWCS'2007), Trondheim, Norway, and at the IEEE Vehicular Technology Conference (VTC'2008 Spring), Singapore.

References

[1] 3GPP TS 36.211 V8.5.0 Release 8, "3rd Generation Partnership Project (3GPP); Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," December 2008.

[2] IST-4-027756 WINNER II, "D6.13.14 WINNER II system concept description," December 2007.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–153, 2001.

[4] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[5] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362–370, 2003.

[6] M. Sternad, T. Svensson, T. Ottosson, A. Ahlen, A. Svensson, and A. Brunstrom, "Towards systems beyond 3G based on adaptive OFDMA transmission," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2432–2455, 2007.

[7] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," November 2007.

[8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[9] I. Ahmad, X. Wei, Y. Sun, and Y.-Q. Zhang, "Video transcoding: an overview of various techniques and research issues," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 793–804, 2005.

[10] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, 2003.

[11] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 122–130, 2006.

[12] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 59–64, 2005.

[13] L.-U. Choi, W. Kellerer, and E. Steinbach, "On cross-layer design for streaming video delivery in multiuser wireless environments," *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, Article ID 60349, 10 pages, 2006.

[14] M. van der Schaar and N. Sai Shankar, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, 2005.

[15] S. Khan, S. Duhovnikov, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for mobile multimedia communication using a common application layer quality metric," in *Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC '06)*, pp. 213–218, Vancouver, Canada, July 2006.

[16] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "A flexible downlink scheduling scheme in cellular packet data systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 2, pp. 568–576, 2006.

[17] X. Zhang, M. Tao, and C. S. Ng, "Time sharing policy in wireless networks for variable rate transmission," in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 4560–4565, Glasgow, Scotland, June 2007.

[18] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," International Telecommunications Union, Geneva, Switzerland, August 1996.

[19] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, "MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication," *Advances in Multimedia*, vol. 2007, Article ID 94918, 11 pages, 2007.

[20] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 1st edition, 2004.

[21] L.-U. Choi, M. T. Ivrlač, E. Steinbach, and J. A. Nossek, "Bottom-up approach to cross-layer design for video transmission over wireless channels," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 5, pp. 3019–3023, Stockholm, Sweden, May-June 2005.

[22] J. Brehmer, C. Guthy, and W. Utschick, "An efficient approximation of the OFDMA outage probability region," in *Proceedings of the 7th Workshop on Signal Processing Advances in Wireless Communications (SPAWC '06)*, pp. 1–5, Cannes, France, July 2006.

[23] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks—the single node case," in *Proceedings of the 11th Annual Conference of the IEEE Computer and Communications Societies (INFOCOM '92)*, vol. 2, pp. 915–924, Florence, Italy, May 1992.

[24] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for: video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[25] L. U. Choi, M. T. Ivrlač, E. Steinbach, and J. A. Nossek, "Sequence-level models for distortion-rate behaviour of compressed video," in *Proceedings of the International Conference on Image Processing (ICIP '05)*, vol. 2, pp. 486–489, Genova, Italy, September 2005.

[26] O. Nemethova, M. Ries, M. Zavodsky, and M. Rupp, "PSNR-based estimation of subjective time-variant video quality for mobiles," in *Proceedings of the International Conference on Measurement of Audio and Video Quality in Networks (MESAQIN '06)*, Prague, Czech Republic, June 2006.

[27] E. Kreyszig, *Advanced Engineering Mathematics*, John Wiley & Sons, New York, NY, USA, 7th edition, 1993.

[28] IST-2003-507581 WINNER, "D5.4 final report on link level and system level channel models, ver. 1.4," November 2005.

[29] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2nd edition, 2002.

[30] W. C. Jakes, *Microwave Mobile Communications*, John Wiley & Sons, New York, NY, USA, 1974.

[31] Joint video team (JVT), "JSVM software manual, version 9.12.2, April 25th, 2008," Heinrich-Hertz-Institut, June 2008, http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm.